BREAST CANCER PREDICTION USING MACHINE LEARNING

Elene Zuroshvili
Central European University
Quellenstrasse 51
Vienna, Austria
Email: zuroshvili_elene@student.ceu.edu

ABSTRACT

This project aims to develop accurate machine learning models that predict cancer recurrence in breast cancer patients based on their clinical information. I compare the performance of the Decision Tree Classifier and the Random Forest Classifier on the "Breast Cancer Data" dataset, which contains 286 instances and 9 attributes, some numeric and some nominal. The objective is to provide a reliable tool for informing clinical decision-making and patient management in oncology.

1 INTRODUCTION

The problem addressed in this machine learning project is to predict the likelihood of cancer recurrence in patients based on various clinical factors. Specifically, I aim to compare the performance of two classification algorithms: Decision tree and Random forest, on this dataset. The dataset[1] used in this project contains information on patients who were previously diagnosed with breast cancer and underwent surgery. The dataset includes columns that describe different clinical features of the patients.

The data set has been previously used in four major researches: Multi-Purpose Incremental Learning System AQ15 and its Testing Application to Three Medical Domains[2], Induction in Noisy Domains[3], Using weighted networks to represent classification knowledge in noisy domains[4], and A Knowledge-Elicitation Tool for Sophisticated Users[5]. The research conducted in these studies involves utilizing machine learning and data mining techniques on a particular dataset, and the resulting findings are analyzed and discussed in conjunction with prior research to determine their potential implications for use in clinical practice.

2 DATA

The data set I am using is called ‘Breast Cancer Data’ and was provided by physicians Matjaz Zwitter & Milan Soklic from the Institute of Oncology, University Medical Center in Ljubljana, and published on 11 July 1988. It is one of three domains provided by the Oncology Institute that has repeatedly appeared in the machine learning literature. This data set includes 201 instances of one class and 85 instances of another class. The instances are described by 9 attributes, some of which are numerical and some are nominal.

2.1 DATA DESCRIPTION

- Number of instances-286
- Number of attributes-9
    - Numerical-1
    - Nominal- 8
- target variable-nominal
    - distribution of target variable: Distribution:

[1] This breast cancer domain was obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Slovenia. Thanks go to M. Zwitter and M. Soklic for providing the data. Please include this citation if you plan to use this database.

[2] Michalski,R.S., Mozetic,I., Hong,J., & Lavrac,N. (1986).

[3] Clark,P. & Niblett,T. (1987)

[4] Tan, M., & Eshelman, L. (1988)

[5] Cestnik,B., Konenenko,I, & Bratko,I. (1987)

1. no-recurrence-events: 201 instances
2. recurrence-events: 85 instances

- Missing values-yes

	class	age	menopause	tumor-size	inv-nodes	node-caps	deg-malig	breast	breast-quad	irradiat
0	no-recurrence-events	30-39	premeno	30-34	0-2	no	3	left	left_low	no
1	no-recurrence-events	40-49	premeno	20-24	0-2	no	2	right	right_up	no
2	no-recurrence-events	40-49	premeno	20-24	0-2	no	2	left	left_low	no
3	no-recurrence-events	60-69	ge40	15-19	0-2	no	2	right	left_up	no
4	no-recurrence-events	40-49	premeno	0-4	0-2	no	2	right	right_low	no

Fig 1: Screenshot of the data set in Python

2.2 DATA UNDERSTANDING

The target variable in my breast cancer machine learning project is a binary variable that indicates whether a patient's cancer has recurred or not. Predicting this variable accurately would be of great benefit to patients, clinicians, and researchers alike. The state of the art in the area of breast cancer recurrence prediction includes various machine learning techniques such as decision trees, and random forests. This project aims to compare the performance of these techniques and select the best one for the task.

The descriptions of the attributes in my dataset are as follows:

Age at diagnosis (continuous variable, 20-90 years), tumor size (continuous variable, 0-250 mm), menopause status ('lt40' for <40 years old, 'ge40' for ≥40 years old, 'premeno' for premenopausal patients), lymph node involvement (discrete variable, 0-40 nodes), tumor location in the breast (categorical variable, 1-4), lymph node involvement (binary variable, 'yes' or 'no'), degree of malignancy (continuous variable, 1-3), breast side ('left' or 'right'), and radiation therapy (binary variable, 'yes' or 'no').

2.3 DATA PROCESSING

I processed the data using Python and Scikit learn. After exploring the data and checking for null values, I removed rows containing question marks to ensure data quality. I then binarized the target variable, with "1" representing "recurrence-events" and "0" representing "no-recurrence-events." I encoded nominal attributes such as breast, node-caps, menopause, breast-quad, and irradiant as numbers and transformed range attributes such as age, tumor-size, and inv-nodes into a single value by calculating their means. The resulting dataset is cleaned and discretized, ready for further analysis.

```
Unique values in column 'class': [0 1]
Unique values in column 'age': [34.5 44.5 64.5 54.5 74.5 24.5]
Unique values in column 'menopause': [2 0 1]
Unique values in column 'tumor-size': [32. 22. 17.  2. 27. 52. 12. 42. 3
7.  7. 47.]
Unique values in column 'inv-nodes': [ 1.  7. 10.  4. 16. 13. 25.]
Unique values in column 'node-caps': [0 1]
Unique values in column 'deg-malig': [3 2 1]
Unique values in column 'breast': [0 1]
Unique values in column 'breast-quad': [1 4 2 3 0]
Unique values in column 'irradiat': [0 1]
```

Fig 2: Unique values of attributes

3 MACHINE LEARNING METHODS

For my project, I chose two different machine learning methods: Decision Tree and Random Forest.t. I chose these two methods specifically, because my data set is labeled, which means I needed to proceed with supervised learning. Most of the attributes in my dataset are categorical, and the target variable is binary, which means that my task is to predict the value of the target variable with the help of attributes.

3. 1 BRIEF DESCRIPTION OF THE METHODS USED

Decision tree is a tree-based model for dataset splitting using features and values as nodes and branches, respectively. Key parameters are criterion, max depth, and min samples to split.

Random forest is an ensemble learning method that uses multiple decision trees with random subsets of features and samples to enhance accuracy and generalization. Important parameters include number of trees, max depth, and min samples to split.

3.2 BRIEF DESCRIPTION OF THE EVALUATION CRITERIA

For my evaluation criteria, I chose several applications: For overall evaluation accuracy, which measures the proportion of correctly classified instances, and for both classes separately-precision, recall, and F1 score which is a harmonic mean of precision and recall that balances both measures. I focused on these metrics specifically, because the distribution of my target variable is not even. No-recurrence-events were seen 201 times, and recurrence events were seen 85 times.

4 EXPERIMENTS

I started the experiment by assigning the target variable and attributes used for the models. I split the dataset into two parts: train set (67 % of the data) and test set (33 % of the data). Before starting with the models, I created a baseline by splitting the data into a training set and a test set. The baseline accuracy was 68.48%. The baseline results for class 1 (recurrence) showed precision, recall, and F1-score of 0.29, 1.00, and 0.45, respectively; however, for class 0 (no recurrence) I got precision, recall, and F1-score of 0.71, 1.00, and 0.83, respectively. These results indicated that the models had room for improvement.

To assess the performance of models, I defined a function that would fit the models to the training data, test them on the test data, and output values for evaluation criteria such as overall accuracy, and for each class precision, recall, and F1-score.

Next, I started to build my model by defining a function that takes as an input the specific data, fits a model to the training set, tests on the test data, and outputs values of evaluation criteria. For the random forest classifier I assigned 10 for estimators, entropy for the criterion, and 5 for the random state. For the decision tree classifier I assigned entropy for the criterion, and 0 for the random state. I proceeded with applying the predefined function to my data.

Classification Report of 'RandomForestClassifier '

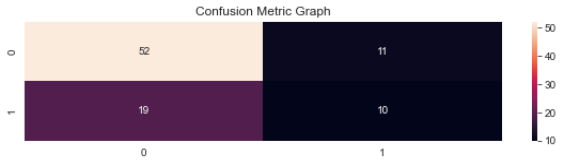
	precision	recall	f1-score	support
0	0.72	0.83	0.77	63
1	0.45	0.31	0.37	29
accuracy			0.66	92
macro avg	0.59	0.57	0.57	92
weighted avg	0.64	0.66	0.64	92

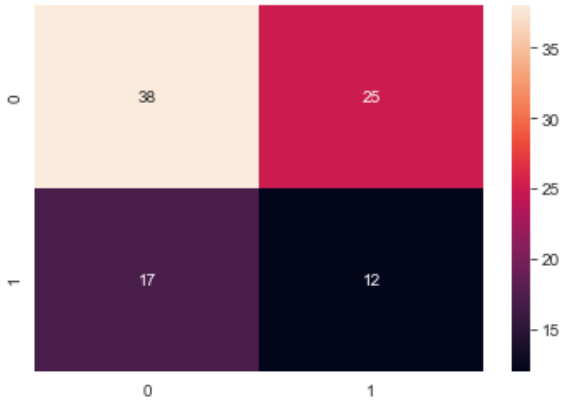
Classification Report of 'DecisionTreeClassifier '

	precision	recall	f1-score	support
0	0.69	0.60	0.64	63
1	0.32	0.41	0.36	29
accuracy			0.54	92
macro avg	0.51	0.51	0.50	92
weighted avg	0.58	0.54	0.56	92

Fig 3: Evaluation scores of train-test split technique

From the classification reports on each algorithm it is clear that the random forest classifier performed with 66.30%, and the decision tree classifier with 54.35%. Both accuracies are less than the baseline: 68.48%. It is also apparent in both cases that both evaluation methods perform better on class 0, than on class 1 if one looks at precision, recall, and F1 scores. However, for both classes the F1 scores are less than the baseline scores-0.83 & 0.45. Before tuning the parameters, I checked the confusion matrices of models to identify whether the model is making more false positives or false negatives.





Figs 4, 5: Confusion matrices of random forest and decision trees with max_depth=5 & max_number_nodes=30 (in this order)

In both matrices it is apparent that true negatives hold the best classification, followed by false positives, false negatives, and the worst classification is true positives. This means that the model is not good at detecting true positives (cancer recurrence) which is a big problem if applied to unseen data. This problem might be the imbalance in the dataset, poor choice of features, overfitting, or underfitting that may need to be improved.

I proceeded with tuning the hyperparameters using the grid search method, imported from Scikit Learn, on my training set. As a result I got the values of best parameters to apply: decision tree-max_features='auto', min_samples_leaf=7, min_samples_split=10; random forest-max_depth=40, max_features='sqrt', min_samples_leaf=2, n_estimators=200. Once I had the parameters in place, I tuned them and used the same function I defined earlier for training and testing the models.

Classification Report of 'RandomForestClassifier '

	precision	recall	f1-score	support
0	0.78	0.94	0.85	63
1	0.75	0.41	0.53	29
accuracy			0.77	92
macro avg	0.76	0.68	0.69	92
weighted avg	0.77	0.77	0.75	92

Classification Report of 'DecisionTreeClassifier '

	precision	recall	f1-score	support
0	0.73	0.90	0.81	63
1	0.57	0.28	0.37	29
accuracy			0.71	92
macro avg	0.65	0.59	0.59	92
weighted avg	0.68	0.71	0.67	92

Fig 6: Evaluation scores of train-test split technique after parameter tuning

All the evaluation metrics, for both classes, increased. The accuracy score for decision trees went from 54% to 71%, and random forest accuracy went from 66% to 77%, both of which now are larger than the baseline:68.48%. F1, precision, and recall, for both classes, increased for both algorithms. However, compared to the baseline, F1 Decision tree metric for both classes is less. These results mean that parameter tuning helped the problem of low criteria. The results were expected, because the evaluation scores are not as high as I aimed them to be, but most are larger compared to the baseline. The limitations of the experiment might be the target variable imbalance, noise and randomness, and bad feature selection; however, if all points are fixed I would expect higher evaluation scores. From what I have at this point, it is interpretable that random forest can classify 77% of the data correctly with support 92 (from 92 instances 77 classified correctly). The decision tree, however, can classify 71 % of the data correctly with 92 support. In addition, it is clear that the models are better at classifying target variable with class 0, than they can class 1 with current hyper parameters. These results are not good, because people with recurrent cancer

will not be classified correctly, which is a major flaw in this experiment.

5 VISUALIZATION

In the project, I used decision trees to make predictions. Decision trees are like a game of 20 questions, where the computer asks questions about different attributes to make a prediction. I made two decision trees, one before pruning and one after pruning. The first tree had a lot of branches and leaves, which made it complicated and hard to understand. It had an accuracy of 54%, which means it correctly predicted 54% of the time.

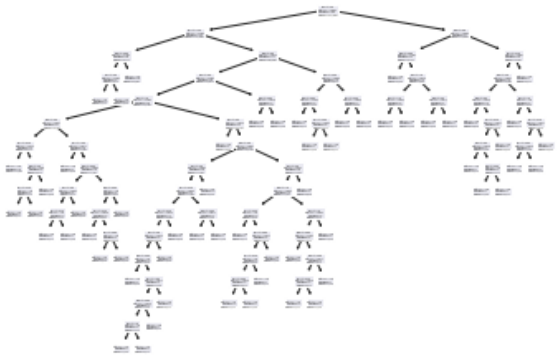


Fig 7: Decision tree before pruning

I then used a technique called pruning to simplify the tree and remove unnecessary branches and leaves. I indicated the same hyperparameters as I got after grid search on the decision tree. The second tree had fewer branches and leaves and was easier to understand. It had an accuracy of 72%, which means it correctly predicted 72% of the time. This improvement in accuracy shows that the pruning technique was successful in simplifying the decision tree and making it more accurate in predicting the recurrence of breast cancer.

Fig 8: decision tree after pruning

6 CONCLUSION

My experiment aimed to predict the likelihood of cancer recurrence in breast cancer patients using machine learning models. I used Decision Tree Model , and Random Forest Model to find the best working parameters to result in high evaluation scores. The findings can contribute to the field of cancer research, after being cleaned of all the limitations and errors, by providing a reliable and accurate tool for predicting cancer recurrence, which can help identify risk factors and develop new treatments. I also found that data mining has several advantages over traditional statistical methods, including the ability to identify non-linear relationships between variables. However, it also has some disadvantages, such as the potential for overfitting and target variable imbalance that might result in inaccurate and overinterpreted results. Regarding future research, I am keen to explore the potential of employing these machine learning models on a larger dataset featuring an increased number of instances and attributes. Moreover, I aspire to delve deeper into the analysis of the influence of each attribute on the target variable and how it contributes to the overall outcome.

References

[1] Michalski,R.S., Mozetic,I., Hong,J., & Lavrac,N. (1986). The Multi-Purpose Incremental Learning System AQ15 and its

Testing Application to Three Medical Domains. In Proceedings of the Fifth National Conference on Artificial Intelligence, 1041-1045, Philadelphia, PA: Morgan Kaufmann.

[2] Clark,P. & Niblett,T. (1987). Induction in Noisy Domains. In Progress in Machine Learning (from the Proceedings of the 2nd European Working Session on Learning), 11-30, Bled, Yugoslavia: Sigma Press.

[3] Tan, M., & Eshelman, L. (1988). Using weighted networks to represent classification knowledge in noisy domains. Proceedings of the Fifth International Conference on Machine Learning, 121-134, Ann Arbor, MI.

[4] Cestnik,G., Konenenko,I, & Bratko,I. (1987). Assistant-86: A Knowledge-Elicitation Tool for Sophisticated Users. In I.Bratko & N.Lavrac (Eds.) Progress in Machine Learning, 31-45, Sigma Press.

[5] Vishabh Goel, Building a Simple Machine Learning Model on Breast Cancer Data, Sep 29, 2018

[6] This breast cancer domain was obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. Thanks go to M. Zwitter and M. Soklic for providing the data.